DOCUMENT RESUME

ED 464 964 TM 033 895

AUTHOR Lee, Yong-Won; Kantor, Robert; Mollaun, Pam
TITLE Score Reliability as an Essential Prerequisite for Validating New Writing and Speaking Tasks for TOEFL.
INSTITUTION Educational Testing Service, Princeton, NJ.
PUB DATE 2002-04-00
NOTE 17p.; Paper presented at the Annual Meeting of Teachers of English to Speakers of Other Languages (Salt Lake City, UT, April 9-13, 2002). For a related paper from the same authors, see TM 033 893.
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS English (Second Language); *Generalizability Theory; *Reliability; *Scores; Second Language Learning; Speech; *Test Construction; Test Items; *Validity; Writing Tests
IDENTIFIERS *Test of English as a Foreign Language

ABSTRACT

This paper reports the results of generalizability theory (G) analyses done for new writing and speaking tasks for the Test of English as a Foreign Language (TOEFL). For writing, a special focus was placed on evaluating the impact on the reliability of the number of raters (or ratings) per essay (one or two) and the number of tasks (one, two, or three). For speaking, the focus was on investigating the impact of the number of tasks (1 though 12) and the number of ratings (1 or 2). Data for the writing study were 488 examinees' responses to 2 reading-writing and 3 listening-writing TOEFL tasks. Data for the speaking study were from 6 listening-speaking, 2 reading-speaking, and 5 independent speaking tasks administered to 261 examinees. By far the greatest source of variation in examinees' test performance was due to differences among test takers' ability as measured by the writing and speaking tasks. This suggests that, as intended, the tasks do distinguish among examinees. Results further suggest that, to maximize score reliability for both speaking and writing, it would be more cost efficient to increase the number of tasks rather than the number of ratings per task. Beyond five or six tasks, there would be a diminishing return for increasing the number of tasks. (Contains 2 tables, 3 figures, and 10 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 964

Score Reliability as an Essential Prerequisite for Validating New Writing and Speaking Tasks for TOEFL

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY

Yong-Won Lee

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Yong-Won Lee

Robert Kantor

and

Pam Mollaun

Educational Testing Service

Rosedale Road Princeton

NJ 08541

BEST COPY AVAILABLE

Paper presented at the annual convention of Teachers of English to Speakers of Other Languages (TESOL), Salt Lake City, Utah, April 2002.

Unpublished Work© 2002 by Educational Testing Service. All Rights Reserved.

TM033895

Introduction

Estimation of score reliability (or score dependability) is usually considered as an essential prerequisite for test validation in language assessment, even though it is not sufficient evidence for validity of score interpretation (Bachman, 1990; Henning, 1987). The investigation of score reliability is basically concerned with what proportion of the total observed score variance among examinees can be attributed to "true score variance" free of measurement error. The "true variance" is also called "universe score variance" in the *geralizability theory* framework or "reliable variance" among some language testers (Oller, 1979). Of course, an empirical test validation process should go beyond computing score reliability and establish relationships between this true score variance and language abilities of interest. Without first making sure of the reliability of test score, however, it may not be hard to establish score validity for language tests, especially for large-scale tests for admissions and selection, such as TOEFL.

Investigation of score reliability takes on a special significance in developing and evaluating new writing and speaking tasks being considered for TOEFL. Both integrated and independent assessment tasks are possible candidates for assessing writing and speaking (Butler, Eignor, Jones, McNamara & Suomi, 2000; Cumming, Kantor, Powers, Santos & Taylor, 2000). These different types of tasks are associated with different types of input stimuli (e.g., auditory texts, reading passages, a stand-alone prompt). Three major types of performance tasks are under consideration to assess writing and speaking: listening-writing (LW), reading-writing (RW), and independent writing (IW) tasks for writing; listening-speaking (LS), reading-speaking (RS), and independent speaking (IS) for speaking. The first two types of tasks in writing are integrated tasks, which require understanding a lecture or a reading, whereas the last one is a traditional TWE-type writing task. In the same vein, the first two types of speaking tasks are integrated task types, while the independent speaking tasks are similar to the ones used in TSE. One might claim that each of the task types in both sections is very different from each other, tapping a somewhat distinct trait, and thus it may be hard to expect a high level of consistency in ranking examinees' performance across different tasks (or task types) in the same section. Adding to the complexity of understanding task reliability is another important source of measurement error: ratings (or raters). When tasks and raters have to be considered as sources of assessment error together, a psychometric framework is necessary that can analyze more than one assessment facet at the same time in the investigation of score reliabilities, such as *generalizability (G) theory*. G- theory provides the most appropriate tool for computing score reliability for such contexts, because it enables us to disentangle differentiated sources of error from the observed score variance.

This presentation reports the results of the generalizability (G) theory analyses that have been done for new writing and speaking tasks. For writing, a special focus was placed on evaluating the impact on the reliability of the number of raters (or ratings) per essay (1 or 2) and the number of tasks (1, 2, or 3). For speaking, the focus was on investigating the impact of the number of tasks (varied from 1 to 12) and of the number of ratings (1 or 2).

Score Reliability in Generalizability (G) Theory

Generalizability (G) theory (Brennan, 1992; Brennan, 2000) is a comprehensive framework that makes it possible to analyze multiple measurement facets simultaneously. Through the *generalizability* (G) and *decision* (D) studies, the *G*-theory enables us to evaluate the relative importance of various sources of measurement error, investigate the impact of various changes in the measurement design (*e.g.*, different numbers of tasks or raters, standardizing rating procedures), and interpret the test score reliability both from norm-referenced and criterion-referenced frameworks (*e.g.*, a *generalizability* coefficient (G or $E\rho^2$) and a *dependability* index (Φ)).

G-theory can be also extended for the multivariate situations where we have a test made up of multiple subsections or subtests and need to investigate the reliability of the composite of subsection scores (Brennan, 2001). Both univariate and multivariate G-theories have been developed within the same G-theory framework, but differ in that the former is suited to analyzing scores from a single test at a time, whereas the latter can also handle scores from multiple subtests simultaneously. For instance, differences between the univariate and multivariate G-theory analyses can be also illustrated in a test developed according to the same "table of specifications". In this type of tests, the same structure of content subsections is applied for assembling multiple forms of the same test. If we are simply interested in the impact on the score reliability of the different numbers of items in the whole test, ignoring the subsection boundary, we can just use the univariate G-theory to estimate variance components and score reliability coefficients for the total test. However, if we are interested in what combination of subtest lengths (or weights) for a fixed total test length can maximize or minimize the composite score reliability of the test, the univariate G-theory analysis does not provide answers to such inquiries, but the multivariate G-theory analysis can.

In the multivariate G- theory design, test subsections or content categories (v) are considered a fixed facet, and the number of elements in each fixed category can be different or unbalanced. It is possible to estimate the variance components for each fixed content category and compute the G and Φ coefficients of the composite scores within the framework of multivariate G-theory. Thus, one clear advantage of multivariate G-theory analysis is that we can examine the impact of different configurations of subsection lengths for the fixed total test lengths on the composite score reliability.

Method

Data

Writing. Data used for study 1 are ratings from a study of new task types. In this study, two Reading-Writing (RW) and three Listening-Writing (LW) tasks were administered to 488 ESL examinees. Examinee responses were rated on a scale of 1 to 5 by two raters from a pool of 27 raters. There were 3 independent writing (IW) tasks used, but each of them was given to three different subgroups of examinees (n_{g1}=162, n_{g2}= 164,

n_{g3} =162). Thus, each subgroup of examinees completed 6 writing tasks (*i.e.*, 5 common integrated tasks +1 independent task specific to each subgroup).

Speaking Section. Data used for this study are also from the study of new task types. Six Listening-Speaking (LS), two Reading-Speaking (RS), and five Independent Speaking (IS) tasks were administered to the examinees. The examinee responses were rated on a scale of 1 to 5 by two independent raters from a pool of 22 raters. Data from only 261 examinees were used for the analysis. Only those examinees who had scores for at least two tasks of all three types were included in the analysis. In addition, two Listening-Speaking tasks based on academic conversations were excluded because it was determined that they did not provide enough substance for spoken responses. Therefore, there were only 11 tasks (4 LS, 2RS, 5IS) included in this study.

Analysis

Writing. The computer program GENOVA (Crick & Brennan, 1983) was used to estimate the variance components for the main and interaction effects and *G* and Φ coefficients for the univariate analysis. Separate analyses were done to estimate the variance components for each of the three subgroups of examinees (n_{g1}=162, n_{g2}= 164, n_{g3} =162) based on the initial data set. The three subgroups were treated as if each of the examinee subgroups had taken 3 different forms of the same test, with LW and RW tasks as common items across all three forms. Each of the variance components was averaged across the three subgroups to obtain stable estimates, and then the G- and Φ coefficients were computed based on these more stable, averaged variance components. Since there was only one task in the IW subsection, it was not possible to do a multivariate G-theory analysis with all three subsections included in a fixed content facet (*v*) for the writing section. There should be at least two tasks in each subsection to estimate variance associated with tasks in the G-study for the subsections.

Speaking. The computer program GENOVA (Crick & Brennan, 1983) was used to estimate the variance components for the main and interaction effects and *G* and Φ coefficients for the univariate analysis. The computer program mGENOVA (Brennan, 1999) was used to conduct the multivariate G-theory analysis.

For the univariate analysis, the same two-facet crossed design ($p \times t \times r'$) as used in writing was employed for the speaking as well to estimate the variance components in the G-study. Multiple D-studies were carried out by varying the number of tasks from 1 to 12 and the number of ratings from 1 to 2. A single facet (p x t) design was used for the averaged rating over two raters to estimate the internal consistency reliability for the different test lengths.

For the multivariate analysis, a two-facet crossed design ($p^{\bullet} \times t^{O} \times r'^{\bullet}$) in the G-study to estimate the variance components for the main and interaction effects for each subsection of task types. It was assumed that persons (*p*) and ratings (r') were crossed with the subsections (*v*) of LS, RS, and IS, but tasks (t) were nested within each subsection (*v*). Multiple D-studies were carried out by varying the number of tasks in

each subsection for the several fixed section lengths. Of particular interest were comparisons of composite score reliabilities for different combinations of subsection lengths for the fixed total test lengths of 3, 4, and 5 tasks. When the total section length was 3, the only possible scenario for representing all three subsections was to take one task for each of the LS, RS, and IS subsections (1-1-1). For the test length of 4, there were three possible scenarios (1-1-2, 1-2-1, and 2-1-1). When the total test length was 5, there were 6 possible scenarios, which included 1-1-3, 1-2-2, 1-3-1, 2-2-1, 3-1-1, 2-1-2. For comparison purposes, two additional combinations for the total test length of 6 were included in the D-study, along with the two longer test scenarios that were used in the G-study to estimate the variance components (4-2-5, 4-3-5).

Results and Discussion

For both the writing and speaking, similar results were obtained about the relative effects of variance components related to tasks and ratings that were estimated through the univariate G-theory analyses. It was found that a single rating scheme had a smaller effect on the score reliability than expected. One striking difference between writing and speaking was that the variance components associated with the task main effect (t) for the speaking was much smaller in proportion (about a fifth) than in writing. Multivariate G-theory analysis also provided useful information to evaluate several possible alternative configurations of tasks for assessing speaking.

Writing

Separate analyses based on three subgroups of examinees and one based on the averaged estimates have yielded similar patterns about the relative proportion of each of the variance components and score reliabilities for various testing conditions. Since an assessment design of particular interest in this study was a test of 3 writing tasks taken by all examinees and rated once or twice, the results are discussed with a focus on these scenarios.

First, the variance component associated with examinees (p) was the largest, which explained about 39% of the total variance estimated for a single observation in the G-study, as shown in Table 1-a. In the single rating situation for 3 tasks, about 65% of the total variance was explained by the person (p) variance, whereas 46% was explained in the double rating situation for a single task. The main effect for ratings (R') was zero, indicating that the ratings were not different in severity between the first and second ratings. However, the main effect for tasks (T) accounted for about 11% of the total variance in the G-study, which may mean that there is a considerable difference in task difficulty among writing tasks.

The examinee-by-task-by-rating interaction (pTR') accounted for the second largest proportion of the total variance for a single observation in the G-study. The pTR' component has explained about 30% of the total variance for a single observation in the G-study, but its relative proportion in the total variance has reduced to 16% for the 3 task scenario with a single rating per essay. The next largest interaction variance component

was the person-by-task interaction (pT) variance component (*i.e.*, 20% for a single observation, 11% of the total variance for the 3-task-and-single-rating scenario). This interaction may indicate some sub-groupings of either tasks or examinees or task-rating combinations that are behaving differently from the rest of the groups. Nevertheless, the examinee-by-rating interaction variance (pR') component was very small (0.5%). It means that examines were rank-ordered consistently the same way overall in the first and second ratings. Lastly, task-by-rating interaction (TR') was almost zero.

Second, due to the fact that the examinee-by-rating interaction (pR') variance was negligible, the impact of one rating per essay was not as serious as expected, as shown in Table 1-b and Figure 1. For instance, the G-coefficient was reduced from about .77 to .69, when the number of ratings was decreased from 2 to 1 for the 3-task situation. Reductions of approximately .09 to .08 in the G-coefficient resulted consistently for each of the 1, 2, and 3 task situations, when the number of ratings was reduced to 1. As the test length increased, the reduced amount in the G-coefficient value became smaller. However, when the number of tasks was increased from 1 to 3 for the single rating scenario, there was a dramatic .25 increase in the G coefficient (*i.e.*, .17 from 1 to 2, .08 from 2 to 3). The similar pattern was observed for the Φ coefficients. It seems that the lost portion of score reliability due to the adoption of the single-rating-per-essay scheme could be probably recovered more than enough by increasing the number of tasks from 1 to 3 for writing.

Finally, as shown in Table 1-b and Figure 1, the line for internal consistency alpha (α) was very close or almost identical to that for the G coefficient for the two-rating-per-essay scenario. This may be attributed to the fact that variance components associated with both the rating main (R') and the person-by-rating interaction (pR') effects were very small, close to zero in the data. Usually, the coefficient α from the p x t design based on the averaged rating should be bigger than the G-coefficient estimated from the p x t x r' design for the two-rating situation, because the random rating error was not incorporated in the α estimate.

Speaking

Similar results were obtained about the impact of single rating on the score reliability both from the univariate and multivariate analyses. Multivariate analysis has additionally provided useful information for comparing various scenarios for the fixed total section lengths. Assessment designs of particular interest were efficient scenarios of 5 speaking tasks taken by all the examinees and rated once or twice in comparison to a liberal scenario of 12 tasks rated twice. The results of the analyses are discussed with a focus on these scenarios.

Univariate Analysis. First, the variance component associated with examinees (p) was the largest, which explained about 51 % of the total variance estimated for a single observation in the G-study, as shown in Table 2-a. In the D-study, about 82% of the total variance was explained by the person (p) variance for the 5-task-and-single-rating scenario, whereas about 93% was explained for the 12-task-and-double-rating situation.

The main effect for ratings (R') was zero, indicating that the ratings were not different in overall severity between the first and second ratings. However, the main effect for tasks (T) accounted for about 2% of the total variance in the G-study, which may mean that there is only a small difference in task difficulty among tasks, especially compared to the writing section.

The examinee-by-task-by-rating interaction (pTR') accounted for the second largest proportion of the total variance for a single observation in the G-study (*i.e.*, about 28% for a single observation). The relative proportion in the total variance of the pTR' component was drastically reduced from 28% to 9%, when the number of tasks was increased to 5 for the single rating scenario. Nevertheless, when the numbers of tasks and ratings were further increased to 12 and 2, respectively, this proportion decreased to 2%, which could be seen as a small change for adding seven more tasks and one more rating per essay to the test. The next largest interaction component was the person-by-task interaction (pT) variance component (*i.e.*, 17% for a single observation, 6% and 3% of the total variance for the above two D-study scenarios). However, the relative effect of the examinee-by-rating interaction variance (pR') component was very small (1.8%), and the task-by-rating (TR') variance was almost zero.

Second, due to the fact that the examinee-by-rating interaction (pR') variance was small, the impact of one rating per essay was not as serious as expected, as shown in Table 2-b and Figure 2. For instance, the G-coefficient was reduced from 0.88 to 0.83 (0.05 reduction) for the 5-speaking-task test when the number of ratings was decreased from 2 to 1. Reduction in G-coefficient values due to the single rating ranged from 0.04 to 0.09 for various section lengths of 1 to 12 tasks. However, when the number of tasks was increased from 1 to 5 for the single-rating scenario, the G-coefficient was increased from 0.52 to 0.83 (a 0.31 increase). There was only a 0.17 additional increase in the G-coefficient (from 0.83 to 0.90), when the numbers of tasks was further increased from 5 to 12. A similar pattern of impact was observed for the Φ coefficients.

Third, as shown in Figure 2, the internal consistency alpha (α) was slightly higher than the G coefficient for the two-rating-per-essay scenario, but very close to the coefficient. This may be attributed to the fact that variance components associated with both the rating main (R') and the person-and-rating interaction effects (pR') were comparatively small for speaking as they were for writing.

Multivariate Analysis. First, the person (p) variance was the largest in each of the three subsections. The main effect for ratings (r') was close to zero in the G-study for all of the three subsections, and the main effect for tasks (t) was also very small, accounting for 1% or less of the total score variances for each subsection in the G-study. It may mean that there is a very small difference in difficulty across tasks even in each subsection. In all three subsections, the second largest variance component was the examinee-task-and-rating interaction (pTR') in the G-study. In the LS and IS subsection, it was followed by the examinee-task (pT) interaction variance in size, but, in the RS subsection, the examinee-by-rating variance (pR') was the third largest variance component followed by the pT variance. Especially, the pR' variance component was significantly smaller in the

LS subsection than in other two subsections, which was partially responsible for the error variance being the smallest and the universe score (or true score) variance being the largest proportionally in the LS subsection.

Second, the correlations between LS and RS subsection scores (.98) and between RS and IS subsection scores (.95) were very high, whereas that between the LS and IS subsection scores was comparatively lower (.89). The very high correlation between the two integrated task types may suggest that tasks in these two subsections may be measuring a very similar construct, even though the stimuli input modes are different. On the other hand, the correlation between the IS and LS subsection scores the lowest among the three subsection pairs (though it can still be seen as high), probably because two task types are comparatively more dissimilar than other pairs in terms of both stimuli input modes (auditory vs. visual) and the nature of tasks (integrated vs. independent).

Third, because the examinee-by-rating interaction (pR') variance was small, the impact of using one rating per task was less serious than expected, as shown in Table 3 and Figure 3. For instance, the G-coefficient for the composite scores was reduced from about 0.85 to 0.78 (.07 reduction) for the 3-speaking task scenario, with one task in each subsection, when the number of ratings was decreased from 2 to 1. For other configurations for the longer section lengths using only one rating reduced G-coefficients by .04 to .06 (See Table 3). A similar pattern of the impact was observed for the Φ coefficients for the composite scores.

Fourth, it seemed that there would be largest gains in composite score reliability, when the number of LS tasks was increased, as shown in Figure 3, because there was slightly less psychometric noise among tasks in the LS subsection. Among the six scenarios for the fixed section length of 5 tasks, the scenario of three LS, one RS, and one IS tasks (3-1-1) seems to achieve the highest G- and Φ coefficients for both one and two rating situations. Similarly for the test length of 4, the highest G- and Φ coefficients were obtained for the 2-1-1 scenario. However, the actual differences in score reliability values among different combinations of subsection lengths for a fixed section length were not seriously large.

Conclusion

By far the greatest source of variation among examinees' test performances was due to differences among test takers' ability as measured by the writing and speaking tasks for the testing conditions of interests. This suggests that, as intended, the tasks do distinguish among examinees. Also, as has been found in numerous other generalizability studies, the next greatest proportion of test variance was attributable to the interaction of person-by-task and person-by-task-by-rater. The variance due to tasks was minimal (particularly for speaking). These latter two findings suggest that, while tasks are on average comparable in difficulty, they are not *uniformly* difficult for all examinees. Finally, the variance due to ratings (or raters) was close to zero, and the examinee-by-rating (rater) was also very small for both writing and speaking.

These results indicate that, to maximize score reliability for both speaking and writing, it would be more cost-efficient to increase the number of tasks rather than number of ratings per task. However, beyond 5 or 6 tasks, there would be clearly a diminishing return for increasing the number of tasks.

Among the three subsections, performance among examinees was slightly more uniform among tasks in the LS (Listening-Speaking) subsection, and, thus, it appears that more gains in composite score reliability could result from increasing the number of LS tasks for the fixed section lengths. However, the actual differences in score reliability values among different combinations of the number of tasks in each subsection for the fixed total section length were quite small. Therefore, the decisions about the possible test configurations should be based on the content and test development considerations as well as the score reliabilities.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: ACT.

Brennan, R. L. (1999). *Manual for mGENOVA Version 2.0*. Iowa City, IA: The University of Iowa.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24 (4)*, 339-353.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL monograph MS -20). Princeton, NJ: ETS.

Cumming, A., Kantor, R., Powers, D., & Santos, T. (2000). TOEFL 2000 writing framework: A working paper (TOEFL monograph MS -19). Princeton, NJ: ETS.

Crick, G. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system (ACT Tech. Bulletin No. 43)*. Iowa City, IA: American College Testing Program.

Henning, G. (1987). *A guide to language testing: Development, evaluation, and research*. Boston, MA: Hienle & Hienle.

Oller, J. W. Jr. (1979). Explaining the reliable variance in tests: the validation problem. In E. J. Biere and F. B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 61-74). Washington, DC: TESOL.

Table 1. Estimated variance components averaged across three examinee subgroups and score reliability coefficients for writing (n_{total} =488; n_1=162, n_2=164, n_3=162)

(a) Variance Components from the G- and D studies

Effects	G-Study		D-Study			
	Single Observation		1 Task, 2 Rating		3 Tasks, 1 Rating	
	Variance	Percent	Variance	Percent	Variance	Percent
Person (*p*)	0.50568	39.07%	0.5057	45.99%	0.5057	65.43%
Task (T)	0.13702	10.59%	0.137	12.46%	0.0457	5.91%
Rating (R')	0	0.00%	0	0.00%	0	0.00%
Person-by-Task (*p*T)	0.26213	20.25%	0.2621	23.84%	0.0874	11.31%
Person-by-Rating (*p*R')	0.00645	0.50%	0.0032	0.29%	0.0065	0.83%
Task-by-Rating (TR')	0.00077	0.06%	0.0004	0.03%	0.0003	0.03%
Person-by-Task-by Rating (*p*TR')	0.3822	29.53%	0.1911	17.38%	0.1274	16.48%
Total	1.29424	100.00%	1.0995	100.00%	0.7728	100.00%

(b) Generalizability (G) coefficients and dependability indices (Φ) based variance components averaged across 3 examinee subgroups

No. of Tasks	One Rating Per Essay		Two Ratings Per Essay		Two Ratings Per Essay
	G	Φ	G	Φ	α
1	0.44	0.39	0.53	0.46	0.53
2	0.61	0.56	0.69	0.63	0.69
3	0.69	0.65	0.77	0.72	0.77
4	0.75	0.71	0.81	0.77	0.82
5	0.79	0.76	0.84	0.81	0.85
6	0.82	0.79	0.86	0.83	0.87

Figure 1. Internal consistency reliabilities (α) for averaged ratings and generalizability coefficients (G or $E\rho^2$) and dependability indices (Φ) for one- and two rating per essay scenarios averaged over 3 subgroups for writing.

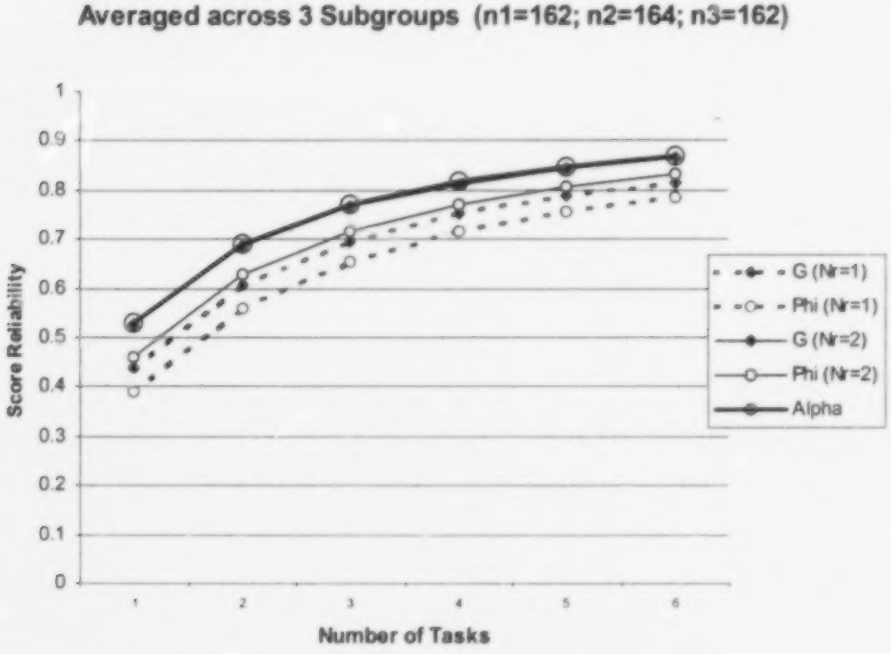

Table 2. Estimated variance components and score reliability coefficients from the G- and D studies for the speaking (n_{total} =261)

(a) Variance Components from the G- and D studies

Effects	G-Study		D-Study			
	Single Observation		5 Task, 1 Rating		12 Tasks, 2 Rating	
	Variance	Percent	Variance	Percent	Variance	Percent
Person (*p*)	0.66856	51.33%	0.66856	82.11%	0.66856	93.36%
Task (T)	0.02184	1.68%	0.00437	0.54%	0.00182	0.25%
Rating (R')	0	0.00%	0	0.00%	0	0.00%
Person-by-Task (*p*T)	0.22484	17.26%	0.04497	5.52%	0.01874	2.62%
Person-by-Rating (*p*R')	0.02367	1.82%	0.02367	2.91%	0.01184	1.65%
Task-by-Rating (TR')	0.00342	0.26%	0.00068	0.08%	0.00014	0.02%
Person-by-Task-by Rating (*p*TR')	0.36007	27.65%	0.07201	8.84%	0.015	2.09%
Total	1.3024	100.00%	0.81426	100.00%	0.7161	100.00%

(b) Generalizability (G) coefficients and dependability indices (Φ)

No. of Tasks	One Rating Per Speech Sample		Two Ratings Per Speech Sample		Two Ratings Per Sample
	G	Φ	G	Φ	α
1	0.52	0.51	0.62	0.60	0.63
2	0.68	0.67	0.76	0.75	0.77
3	0.75	0.75	0.82	0.81	0.83
4	0.80	0.79	0.86	0.85	0.87
5	0.83	0.82	0.88	0.87	0.89
6	0.85	0.84	0.89	0.89	0.91
7	0.86	0.86	0.91	0.90	0.92
8	0.87	0.87	0.91	0.91	0.93
9	0.88	0.88	0.92	0.92	0.94
10	0.89	0.89	0.93	0.92	0.94
11	0.90	0.89	0.93	0.93	0.95
12	0.90	0.90	0.94	0.93	0.95

Figure 2. Internal consistency reliabilities (α) for averaged ratings, generalizability coefficients (G) and dependability indices (Φ) for one- and two rating per essay scenarios for different section lengths for speaking

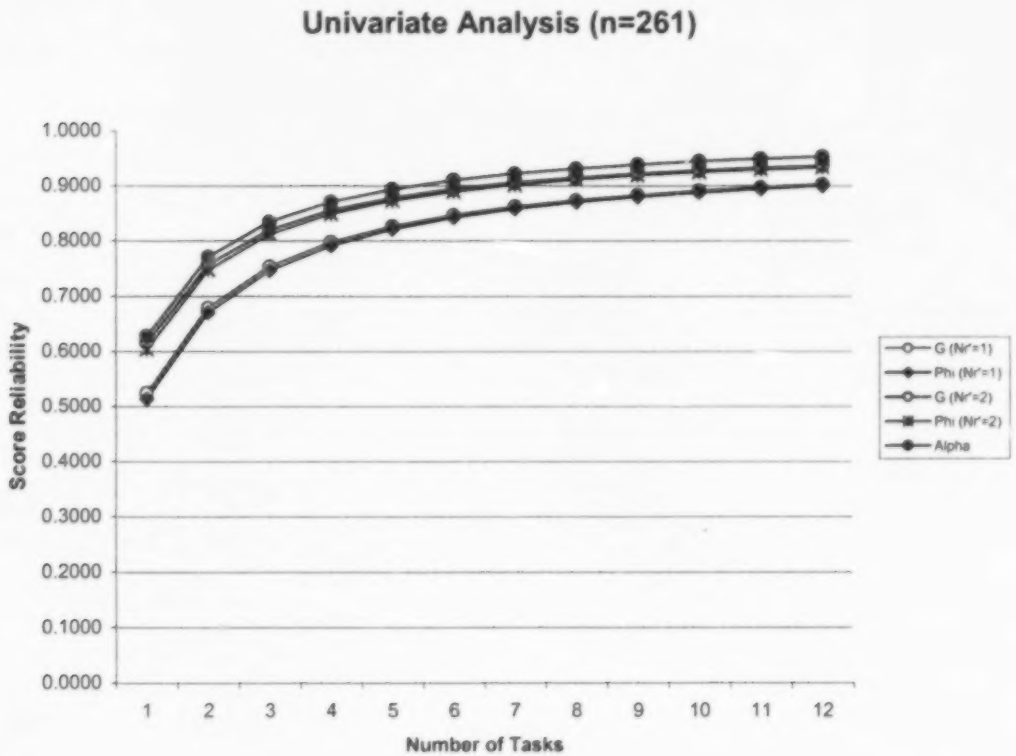

Table 3. Generalizability (G) and dependability indices (Φ) for composite scores for different combinations of subsection lengths for fixed total section lengths (n_{total} =261) for speaking

D-Study Scenarios				Single Rating		Double Rating	
Total No. of Tasks	No. of LS Tasks	No. of RS Tasks	No. of IS Tasks	G	Φ	G	Φ
3	1	1	1	0.78	0.78	0.85	0.84
4	2	1	1	0.83	0.82	0.88	0.88
4	1	2	1	0.81	0.81	0.87	0.87
4	1	1	2	0.81	0.81	0.87	0.87
5	3	1	1	0.85	0.85	0.90	0.90
5	2	2	1	0.84	0.84	0.90	0.89
5	2	1	2	0.85	0.84	0.89	0.89
5	1	2	2	0.83	0.83	0.89	0.89
5	1	1	3	0.83	0.83	0.88	0.88
5	1	3	1	0.82	0.82	0.88	0.88
6	2	1	3	0.86	0.85	0.90	0.90
6	1	1	4	0.84	0.84	0.89	0.89
11	4	2	5	0.90	0.90	0.94	0.94
12	4	3	5	0.91	0.90	0.94	0.94

Figure 3. Generalizability coefficients (G) and dependability indices (Φ) for composite scores for one- and two rating per essay scenarios for different combinations of subsection lengths for fixed total section lengths for speaking